EXHIBIT E

IN THE UNITED STATES DISTRICT COURT
DISTRICT OF DELAWARE

AFFYMETRIX, INC., a Delaware corporation,	)	
Plaintiff,	)	No. C4-901 JJF
v.	)	
ILLUMINA, INC., a Delaware corporation,	)	CERTIFIED COPY
Defendant.	)	

* * * * * * * * * * * * * * * *

THE VIDEO DEPOSITION OF KENNETH L. BEATTIE, Ph.D.,

February 1, 2006

==

TRUESDEL & RUSK
Registered Professional Reporters

David J. Doyle Court Reporter
8321 Block House Way, Apartment 632
Knoxville, Tennessee 37923
(248) 703-7419
DD860@AOL.COM

Q. Now, what do you --

A. Then I did a post doc after that at Yale.

Q. And Dr. Beattie, just to help out the court reporter --

A. Okay.

Q. -- just pause a little bit after my questions --

A. Okay.

Q. -- and then go ahead and answer, okay?

So you were telling us that you did a post doc at Yale?

A. Yes.

Q. When was that?

A. From '74 to '77. It was in Yale Medical School, Charles Reading's group working on entomology of DNA.

Q. What did you do after your post doc?

A. Then I looked for jobs. And I got interviewed at several places. And I chose to accept a job in the biochemistry department at Baylor College of Medicine in Houston. And I moved there November of 1977.

Q. And did you hold a specific title at

Baylor?

A. Assistant professor at first, and then associate professor later on.

Q. Do you remember when you became associate professor?

A. I don't remember the specific date. I can't tell you.

Q. Oh, no, that's fine. That's fine.

A. Maybe five years later or something.

Q. Okay. What did you -- how long were you at Baylor?

A. I was at Baylor until -- it is a little complicated because I was in the biochemistry department for maybe fifteen years.

And then I moved to another department called the Center for Biotechnology in the Woodlands. It's a little bit out on the north side of Houston. And that was about 1990. And I was -- kept a faculty position there while I was working in my lab at a place called Houston Advanced Research Center or ARC as we called it. It's a non-profit research organization. It's in the Woodlands right -- a few blocks away from the center for biotechnology. And so I was there about five years.

And then I moved back to Oak Ridge

National Lab and stayed there for five years as a staff scientist until about -- let's see, that would have been about four years ago, 2002 or 2001, perhaps.

Q. Okay. So Dr. Beattie, when you were in Houston, did you hold any other positions or titles?

A. Well, I was -- the research I was doing in my lab used a lot of synthetic DNA oligonucleotides. And at that time you couldn't really buy them. There weren't any vendors.

But there was some equipment you could buy to synthesize them. And we purchased a small -- not -- a semi-automated system from a company called from Cruachem in Scotland and we started making oligonucleotides. And in my grant proposals I proposed to do a lot of site directed mutagenesis. And it required a lot of synthetic DNA, more than what we could really afford in the budget.

So my technician and I started working with a new process for making synthetic oligonucleotides, what we called a segmented synthesis approach. And I can explain that more if you would like.

But it allowed us to make large number

A. Yes.

(Reporter labeling previously marked Exhibit 260).

BY MS. TANG:

Q. Dr. Beattie, the court reporter has just handed you what was previously marked as Exhibit 260 bearing production numbers IAFP 597861 to 882 entitled Wolf Trap Genome Sequencing Conference poster abstracts.

If you could just -- and I'm not asking you to read through the whole thing, just take a look at it and let me know if you have ever seen this document before.

A. Yes. This would be the abstracts of the meeting that was supplied to all the participants, along with the program that's in the front.

Q. So was the Wolf Trap conference a public conference?

A. It was public in the sense of anyone could go there if they had relevant research. I think they may need to be applied like a lot of meetings, and accepted and come to the meeting. Scientists from all over the world.

Q. Would you agree that the information

disclosed during the meeting was intended to be information that could be used by others?

MR. FAHNESTOCK: Objection, leading.

THE WITNESS: In the sense of standard dissemination of scientific results; in that sense, yes.

BY MS. TANG:

Q. And here I'm thinking of say there was poster presentations, right?

A. Yes.

Q. And there were talks, correct?

A. Yes.

Q. And it was expected that people would take notes at these talks --

MR. FAHNESTOCK: Objection, leading.

BY MS. TANG:

Q. -- during poster presentations, right?

A. I would expect that most people would take notes, yes, if it's a topic that they are especially interested in.

Q. I'd like you to take a look at Exhibit 260 at IAFP 597869. If you look at the bottom right-hand corner. Look at 869.

A. Okay.

Q. And in particular I'm looking at abstract number six.

A. Yes.

Q. If you could read abstract number six. And I apologize, it's very tiny print. Let me know when you are done reading that.

A. Okay.

VIDEO OPERATOR: Pardon me while we pause. I am going to change the tape. We will go off the record. The time is 10:22.

(Interruption).

VIDEO OPERATOR: The time now is 10:23. We are back on the record.

THE WITNESS: Okay.

BY MS. TANG:

Q. And I just want to make clear, you mentioned when you attended the conference you were given this book of abstracts, correct?

MR. FAHNESTOCK: Objection, leading.

THE WITNESS: That would be the standard procedure.

BY MS. TANG:

Q. So when you say "standard procedure," each of the attendees would receive a book of abstracts?

MR. FAHNESTOCK: Objection, leading.

THE WITNESS: That's true. We all received the abstracts.

BY MS. TANG:

Q. And if an abstract appeared in the book, what did that mean typically?

A. It would mean that there was either a poster presentation or an oral presentation on that topic describing what was summarized in the abstract.

Q. Do you recall if Dr. Crkvenjakov and Drmanac either gave a talk or presented a poster at this Wolf Trap meeting?

A. I believe they presented two posters. And I recall one of them was on general SbH, including the forward or the original version of it where you array the DNA samples and then hybridize over and over with oligonucleotides.

And the other one I recall seeing was this discrete particle approach.

Q. So let's focus in on this particular abstract which -- can you tell which approach is being discussed here?

A. This would be the discrete particle approach which involves the use of these coded beads.

Q. And why do you say that you can tell

the discussion has to do with coded beads?

A. Because here they are mentioning the fact that you would need to decode the beads either by some physical characteristic or by hybridization over and over again with different oligonucleotides that you could see which ones were contained on a specific bead.

Q. So do they mention different physical characteristics that you could use to mark the beads?

A. I believe it mentions several in the abstract here, color, size, shape or other ways besides the use of the oligonucleotide signature.

Q. So essentially in this abstract Dr. Drmanac and Crkvenjakov have disclosed several different ways of encoding a bead so you can figure out what oligos attach to the bead, is that correct?

A. Yes.

MR. FAHNESTOCK: Objection to leading.

BY MS. TANG:

Q. And you mentioned that they also tell you how to decode the beads, is that right?

A. Yes.

MR. FAHNESTOCK: Objection, leading.

bead.

Q. So is it accurate to say that there was -- on the bead there were essentially two types of oligos, there was the probe oligo that you were interrogating the sequence of?

A. Yes.

Q. And there was also a marking oligo to be used for decoding the bead?

A. Yes.

MR. FAHNESTOCK: Objection, leading.

THE WITNESS: And it would have to be shorter than the probe one because when you then take the fluorescently tagged sample, DNA of unknown sequence and hybridize, you want it to only hybridize to the longer probes and not the short ones that were used for decoding, otherwise everything would light up.

BY MS. TANG:

Q. Does this abstract tell you what size of oligo probe that doctors Drmanac and Crkvenjakov were thinking of? And by "oligo probe" I am not talking about the marking probes, but I mean the marker oligos, but rather the ones that you are using to interrogate the sequence.

MR. FAHNESTOCK: Objection to form.

THE WITNESS: It tells what the -- it mentions the 12 to 15 base probe, 15 base long probe to be interrogating the sequence.

BY MS. TANG:

Q. So I just want make sure that I understand everything that we have talked about now about this abstract. So I'm going to ask you kind of a series of questions and you let me know if I'm correct or not.

A. Okay.

Q. Okay? This abstract is to do with inverse SbH?

A. Yes. Well, it actually refers to both. You could also do it in a traditional or the initial way.

Q. The abstract discusses using one times ten to the seventh to one times ten to the ninth different discrete particles?

A. Yes.

MR. FAHNESTOCK: Objection, leading.

BY MS. TANG:

Q. And will you understand me if I say beads?

A. Yes.

Q. So each of the beads carries a

unique 12 to 15 mer, m-e-r?
A. Yes.
Q. The beads also carry unique combinations of marking oligos?
A. Yes.
Q. And those marking oligos are essentially an encoding system for the beads, correct?
A. Yes.
MR. FAHNESTOCK: Objection to form.
THE WITNESS: Encoding and decoding.
BY MS. TANG:
Q. And this collection of one times ten to the seventh different beads, that's a collection of beads essentially?
A. Yes.
Q. And the beads contain binding polymers with different target specific oligos attached?
A. Yes.
Q. And those target specific oligos are the 12 to 15 mers?
A. Yes.
MR. FAHNESTOCK: Objection to form.
BY MS. TANG:

Q. The abstract also mentions that you would want to do 40 hybridizations with marker oligos to define the associations?

A. Yes.

Q. This is a decoding step so you can identify the location of a particular oligo on the bead?

A. That's right.

Q. And the discrete particles or beads are on monolayer in this abstract?

A. Yes.

Q. Would you agree that the slide has at least a thousand different beads with different oligos attached?

MR. FAHNESTOCK: Objection to form.

THE WITNESS: Yes. I would think it would contain -- it would need to contain the full. I think they said ten to the seventh to ten to the ninth and probably more, a lot more --

BY MS. TANG:

Q. Do you have an understanding --

A. -- of redundancy.

Q. Do you have an understanding as to the density of the beads on the slide?

A. I don't think it specifically

Q. The old dot matrix printer?

A. Yes. They brought with them printed sheets and lined them up into a title.

Q. Do you know -- you say that you recognize this document. What is it?

A. This is -- this is the abstract -- I mean, I'm sorry, the poster presentation that corresponds to this abstract on the previous Exhibit 260.

Q. So just so we're clear for the record, you recognize Exhibit 275 as the poster that Dr. Crkvenjakov and Drmanac presented at the Wolf Trap genome sequencing conference in 1989?

A. Yes.

MR. FAHNESTOCK: Objection, leading.

BY MS. TANG:

Q. I'd like you to turn to the page marked IAFP 598104. Look at the last three numbers, 104 --

A. Okay.

Q. -- of Exhibit 275. I want to draw your attention to the middle section that's labeled for inverse SbH.

A. Yes.

Q. Can you read through that whole

Q. So with this understanding of the limit of sensitivity, would you agree that it would be possible to identify one bead with fluorescently labeled DNA attached to it?

A. I think it would be feasible. I mean you would want to do experiments to see what the level of the actual sensitivity could be achieved.

Q. Now, you mentioned earlier that you recognized Exhibit 275 the poster that Dr. Drmanac and Crkvenjakov presented at the Wolf Trap meeting in 1989, correct?

A. Yes.

Q. Do you recall where the posters were laid out?

A. They were in an area adjacent to the meeting room where there was sort of a large room with all the seats and the projection. And my recollection was out in the sort of the in a nearby area. I don't know if it was all one room or hallways or combined. They were the posters that people would circulate through and read them.

Q. So the attendees had access to all the posters that were up?

A. Oh, yes.

MR. FAHNESTOCK: Objection, leading.

A. Okay.

Shall I read this one, too?

Q. Yes, if you could.

A. Let's see, this one is December 7th. This one is also December 7th. Okay.

Okay. There is a lot there.

Q. There certainly is. But we will take things one at a time.

Let's start on Exhibit 276 at the last page IAFP 598053 -- it is actually on the other side. There you go.

So this appears to be a telefax from you to Dr. Crkvenjakov, is that correct?

A. Yes.

Q. And you are attaching looks like four pages including the cover?

A. Yes.

Q. Okay. So let's look at the first thing that you have attached. And this would be Exhibit 276 at IAFP 598052. I'm going backwards.

A. Okay.

Q. Yes. There we go. And this is a letter from yourself to Dr. Bruce Jacobson, is that correct?

MR. FAHNESTOCK: Objection to form.

THE WITNESS: I'm sorry. I think I'm looking at the wrong one.

BY MS. TANG:

Q. Oh, this letter -- oh, I'm sorry.

A. It's to me from Bruce Jacobson.

Q. There you go.

A. Yes. At Oak Ridge.

Q. Can you tell me what is Dr. Jacobson discussing in this letter?

A. He is discussing a couple of grant applications that they are intending to submit in the very near future and to do two types of SbH projects that they would like to do at Oak Ridge.

One is -- refers to a chip -- Bob Foote. And it doesn't tell too much about it in this letter. But the other one is a bead technology that's similar to the DP concept. F-o-o-t-e. He was a scientist at Oak Ridge at that time and still there I think.

Q. Okay. Let's talk about this proposal that they have on the second proposal on beads.

You said that was similar to, I'm sorry, similar to what?

A. Similar to the discrete particle chip that was described here, similar in concept, but

it would be performed differently, not on a surface, but in a flowing stream of little droplets that could then decode them very quickly with the optical systems they had at Oak Ridge.

Q. Okay. So when you say "similar," you are talking about a collection of beads?

MR. FAHNESTOCK: Objection to form.

THE WITNESS: Yes, it would be a collection of beads.

BY MS. TANG:

Q. And you are talking about beads with binding polymers of different target specific sequences attached?

MR. FAHNESTOCK: Objection to form.

THE WITNESS: That would be my understanding, although, I think I would need to read through it again.

BY MS. TANG:

Q. Do you want to read through it again?

A. All the details.

Q. Go ahead and read through it again. I think it's just the -- yes.

A. Okay. Let me look again.

Okay.

Q. And I think -- let me refer to the portion that says, "we propose to label beads with three stable isotopes."

Do you see that part?

A. Yes.

Q. And then do you see that it then goes on to say, each bead would have a unique oligonucleotide attached?

A. Yes.

Q. So is it your understanding that they are proposing to have a collection of beads?

A. Yes.

Q. And do you understand that this collection of beads each would have a unique DNA sequence attached?

MR. FAHNESTOCK: Objection to form.

THE WITNESS: Yes. There is a little bit of confusion because he first says three stable isotopes then immediately says five, five stable isotopes that could be in different combinations to give a signature of at least ten to the fourth combinations, ten thousand combinations.

BY MS. TANG:

Q. Which is he talking about when you say three versus the five?

MR. FAHNESTOCK: Objection to form.

THE WITNESS: Oh, that's the coding and decoding method. So each bead would have three and it looks like they would propose five different discrete identifiable isotopic tags. So they would attach them to the beads and different mixtures, different known. So one would have the different -- they would have different mixtures of five -- oh, maybe it's three out of five. Maybe they would have three different ones and another one have another set of three. I forget. You have to think about it some more. But that would be the coding and decoding mechanism equivalent to the color or shape that's -- was described in there in their DP description.

BY MS. TANG:

Q. So it is your understanding that what Dr. Jacobson is proposing here is an encoding system for the collection of beads?

MR. FAHNESTOCK: Objection, leading.

THE WITNESS: Yes.

BY MS. TANG:

Q. And the purpose of the encoding system is to be able to identify the unique oligo attached to each bead?

MR. FAHNESTOCK: Objection to form.

THE WITNESS: Yes. I don't think it's fully described in this letter, but you can think about it and come up with a scheme where you have the five labels and have different mixtures of them containing the different discernable isotopes in different ratios and whatnot coming up with signatures that could be ten thousand different combinations, could be coding and then decoding each bead.

BY MS. TANG:

Q. Now, why, can you explain to me your understanding of this sentence. It's the third paragraph from the top or the second paragraph from the bottom.

In particular I'm looking at the last sentence of the two sentence paragraph. It begins, "I told them we would consider any bead project a collaborative one."

A. Yes.

Q. What is your understanding of that statement?

A. They must have had some prior discussions with them, with the -- Crkvenjakov about whose -- which kind of projects are in whose territory.

And the beads, like for example, the what was in the abstract from the Wolf Trap meeting was clearly a concept being worked on in and owned by the Crkvenjakov group. I think that's what that is referring to.

Q. The point, though, is that Dr. Jacobson in this letter is proposing a different approach, as you said, right?

A. Hm-hmm. Yes. That would be meaning they want to use beads, but in a different way. There is some common features of the process, but they are going to be implemented in a completely different way.

Q. And can you tell me --

A. So it would be more of a collaboration instead of a -- not an independent one either.

Q. In Dr. Jacobson's proposal is the detection system the same as in Dr. Drmanac's and Crkvenjakov's?

MR. FAHNESTOCK: Objection to form.

THE WITNESS: No, no. In this bead technology, that they are describing here to be done at Oak Ridge with Mike Ramsey, they're going to use RIS, which is resonance ionization spectroscopy.

EXHIBIT F

WOLF TRAP GENOME SEQUENCING CONFERENCE

October 24th-26th, 1989

~~POSTER ABSTRACTS~~

Washington D.C.

for poster 6, 27

1) SEQUENCING DNA WITH STABLE ISOTOPES

K.B. Jacobson, F.W. Larimer, R.P. Woychik, R.S. Foote, G.M. Brown, K. L. Burchett, D.A. Jacobson, R.A. Sachleben, F.V. Sloop, H.F. Arlinghaus and N. Thonnard. Biology, Chemistry, Health Safety Research Divisions, Oak Ridge National Laboratory and Atom Sciences, Inc.

2) TOWARDS SEQUENCING LONGER FRAGMENTS BY DNA TRAPPING ELECTROPHORESIS

Levy Ulanovsky, Guy Drouin and Walter Gilbert. Harvard University.

3) MULTIPLEX DNA SEQUENCING WITH SANGER CHEMISTRY AND CHEMILUMINESCENT DETECTION

Stephan Beck, Louis D'Angio, Theresa S. Dunne, Richard P. Hodge, Carol Kissinger, Thomas O'Keeffe, Thomas Templeman and Hubert Koster. MilliGen/Biosearch.

4) NEW SYNCHRONOUS LUMINESCENCE TECHNIQUE FOR IMPROVED DNA SEQUENCING

T. Vo-Dinh, R.W. Johnson and H.F. Cheng. Advanced Monitoring Development Group, Health and Safety Research Division, Oak Ridge National Laboratory.

5) IMPROVED SEQUENCING OF DOUBLE-STRAND PLASMID TEMPLATES WITH MODIFIED TAQ DNA POLYMERASE

David Thompson, Tom Burke and Sandra Spurgeon. Promega Corporation.

6) MINIATURIZATION OF SEQUENCING BY HYBRIDIZATION (SBH): A NOVEL METHOD FOR GENOME SEQUENCING

R. Crkvenjakov, R. Drmanac, Z. Strezoska and I. Labat. Genetic Engineering Center.

7) DIRECT SEQUENCING OF AMPLIFIED DNA FROM HIV INFECTED INDIVIDUALS

D. Amorese, T. Norris, M. Kirsch and K. Adler. Medical Products Department, E. I. du Pont de Nemours & Co.

8) TAQ THERMAL CYCLING FOR DIDEOXY SEQUENCING

Sandy M. Koepf, Leslie Johnston-Dow, James A. Fisher, Marianne Hane, Lincoln J. McBride and Cheryl Heiner. Applied Biosystems, Inc.

9) OPTIMIZATION OF ASYMMETRIC POLYMERASE CHAIN REACTION FOR RAPID FLUORESCENT DNA SEQUENCING

Chia Chen, Richard K. Wilson, Ben Koop and Leroy Hood. Division of Biology, California Institute of Technology.

10) AUTOMATED APPROACHES TO DNA SEQUENCING

Bruce A. Roe, Elaine Mardis and Stephanie Pyle. Department of Chemistry and Biochemistry, University of Oklahoma.

11) AUTOMATION OF THE EXONUCLEASE III DIGESTION REACTION USING A BIOMEK 1000 AUTOMATED LABORATORY WORKSTATION

Anthony R. Kerlavage and J. Craig Venter. Section of Receptor Biochemistry and Molecular Biology, LMCN, NINDS, NIH.

12) RAPID DNA SEQUENCING USING THE MULTIWELL MICROTITRE PLATE SEQUENCING SYSTEM CONTAINING LYOPHILISED REAGENTS

Peter J. Ellis, Ian R. Felix, Alison C. Sweet, Ronald H. Jackson and Jeremy N.B. Walker. Amersham International.

13) THE USE OF AUTOMATED, PREPARATIVE ELECTROPHORESIS TO FACILITATE SUBCLONING OF FRAGMENTS FOR SEQUENCE ANALYSIS

W. Richard McCombie, David Iovannisci, J. William Efcavitch and J. Craig Venter. Section of Receptor Biochemistry and Molecular Biology, NINDS, NIH and Applied Biosystems, Inc.

14) AN AUTOMATED DNA SEQUENCE FILM READER

Steve Ferris, Bio-Rad Laboratories.

15) A DNA SEQUENCE FILM READING WORKSTATION

Murray R. Summers, Mike Kelly, Abe Coriat, Ilya Ravkin and Al Delfino. IntelliGenetics, Inc. and Applied Imaging, Inc.

16) RAPID TEMPLATE PURIFICATION METHODS FOR DNA SEQUENCING: CONVENTIONAL AND AUTOMATED DNA SEQUENCING OF SEPHAGLAS™ M13 MINIPREPS

S. Duthie, D. Boyer, B. Walsh, T. Pham and H. Osterman. Pharmacia P-L Biochemicals Inc.

17) EFFICIENCY CHARACTERIZATION OF A FLUORESCENCE BASED AUTOMATED DNA SEQUENCER

F. H. Fraser, B. K. Burnett, J. P. Mizzer, A. M. Hochberg and J. W. Eveleigh. E. I. Du Pont de Nemours & Co., Inc.

18) FLUORESCENT DNA SEQUENCING USING DYE-LABELED NUCLEOTIDE TERMINATORS

Vergine Chakerian, Charles R. Connell, Steven Fung, Davis Hershey, Linda G. Lee and Sam L. Woo. Applied Biosystems, Inc.

19) TOWARDS LARGE SCALE DNA SEQUENCING

John A. Brumbaugh, Lyle R. Middendorf, Daniel L. Grone and Jerry L. Ruth. School of Biological Sciences, University of Nebraska, LICOR, Inc., and Molecular Biosystems.

20) PHAGE T4 IN VITRO DNA PACKAGING SYSTEM FOR CLONING LONG DNA MOLECULES

V. Basaveswara Rao and Lindsay W. Black. Department of Biological Chemistry, Univ. MD. Medical School.

21) GENOMIC SEQUENCE OF A 51 BASE PAIR REPEAT GENE OF PLASMODIUM FALCIPARUM: IMPLICATION OF SEQUENCING REPETITIVE DNA SEQUENCE OF HUMAN GENOME

Jingdong Zhu and Michael R. Hollingdale. Biomedical Research Institute.

22) RAPID COSMID FINGERPRINTING AND CONTIG ASSEMBLY

E. Branscomb, P. de Jong and A.V. Carrano. Biomedical Science Division, Lawrence Livermore National Laboratory.

23) METHOD AND STRATEGY FOR DIRECTLY SEQUENCING COSMID INSERTS: SEQUENCING THE HUMAN T-CELL RECEPTOR ß-GENE FAMILY

Jerry L. Slightom, David R. Siemieniak, Leang C. Sieu, Leroy Hood and Morris Goodman. Molecular Biology Unit 7242, The Upjohn Company; Division of Biology, California Institute of Technology and Department of Anatomy and Cell Biology, Wayne State University.

24) NUCLEOTIDE SEQUENCE DETERMINATION OF THE GALAGO 45 KB β-TYPE GLOBIN GENE CLUSTER

Danilo A. Tagle, Philip Benson, David Fitch, Michael Stanhope, Jerry L. Slightom and Morris Goodman. Depts. of Molecular Biology & Genetics, Anatomy, Wayne State and Div. of Molecular Biology, The Upjohn Company.

25) BOVINE IMMUNODEFICIENCY-LIKE VIRUS: NUCLEOTIDE SEQUENCE ANALYSIS AND EVOLUTIONARY COMPARISON TO OTHER RETROVIRUSES

Kevin J. Garvey, M. Steven Oberste and Matthew A. Gonda. Laboratory of Cell and Molecular Structure, Program Resources, Inc., National Cancer Institute-Frederick Cancer Research Facility.

26) SOFTWARE FOR TRANSFER AND GRAPHICAL EDITING OF DATA FROM AUTOMATED DNA SEQUENCERS.

James Stein, Janice McCombie, Robert Jones, Jeannine Gocayne, W. Richard McCombie, Anthony Kerlavage and J. Craig Venter. Applied Biosystems, Inc. and Section of Receptor Biochemistry and Molecular Biology, NINDS, NIH.

27) ALGORITHM FOR SEQUENCE GENERATION FROM K-TUPLE WORD CONTENT

R. Drmanac, I. Labat and R. Crkvenjakov. Genetic Engineering Center, Belgrade.

28) THE USE OF ARTIFICIAL INTELLIGENCE METHODS IN AUTOMATED DNA SEQUENCE ANALYSIS SOFTWARE

T. Allen, C. Shearer-Cooper and G. Page. EG&G Biomolecular.

29) GENLANG: A COMPUTATIONAL LINGUISTIC SYSTEM FOR SEQUENCE DESCRIPTION, SEARCH AND ANALYSIS

Michiel O. Noordewier and David B. Searls. Computer Science Department, University of Wisconsin and Paoli Research Center, Unisys Corporation.

30) DNA EXON BOUNDARIES DETECTED BY EVOLUTIONARY CONSERVATION

G. William Moore, Stephen Williams, Jerry Slightom, Danilo Tagle, David Fitch and Morris Goodman. Baltimore Veterans' Administration Medical Center and Wayne State University School of Medicine.

31) ESSENTIAL ROLE OF THE PROTEIN SEQUENCE DATABASE IN GENOME ANALYSIS

W.C. Barker, D.G. George and L.T. Hunt. Protein Identification Resource, National Biomedical Research Foundation.

32) DEVELOPING METHODOLOGY FOR CLONE CHARACTERIZATION IN A REPOSITORY OF DNA CLONES

J.A. Aebig, A.S. Durkin, D.L. Perino, J. Tomlin, D. Maglott and W.C. Nierman. American Type Culture Collection.

1. Sequencing DNA with Stable Isotopes

K. B. Jacobson,[1] F. W. Larimer,[1] R. P. Woychik,[1] R. S. Foote,[1] G. M. Brown,[2] K. L. Burchett,[2] D. A. Jacobson,[3] R. A. Sachleben,[2] F. V. Sloop,[1] H. F. Arlinghaus,[4] and N. Thonnard,[4] [1]Biology, [2]Chemistry, [3]Health Safety Research Divisions, Oak Ridge National Laboratory, and [4]Atom Sciences, Inc., Oak Ridge, TN 37831

A rapid DNA sequencing procedure is being developed that will use resonance ionization mass spectrometry (RIS) to detect stable isotopes that label the DNA. Isotopes of iron, tin and other elements with multiple, stable isotopes are incorporated into stable organometallic complexes that in turn are covalently attached to the 5' end of oligonucleotides. After the Sanger or Church procedure the isotope will be located in the sequencing ladder, that is on a nylon membrane, by scanning each lane with the RIS. The scanning beam of the RIS can be focused from 1 μm to several hundred μm diameter so the resolution and sensitivity can vary interdependently. Two ways of obtaining high density of information from a sequencing gel are: 1) operate at high resolution using four isotopes to label the A, G, C and T terminated fragments (Sanger) and run them in a single gel lane; 2) label 40 DNA probes (Church multiplex method) with 40 isotopes and hybridize all electrophoretically separated fragments from 40 DNAs simultaneously. In either case the RIS scan will locate all the individual isotopes of one or two elements in each scan and thus provide the location of the DNA fragments. Organometallic chemistry has been developed to incorporate iron and tin isotopes into oligonucleotides. The RIS has detected 0.02 fmole of ^{57}Fe in ferrocene-labeled DNA and has located iron-labeled oligonucleotides on the nylon membrane; commercial Gene Screen has iron contamination that interferes with sensitivity studies. Studies on Sn-labeled oligonucleotides are also underway. (Research sponsored by the OHER, U.S. DOE, under Contract DE-AC05-84021400 with the Martin Marietta Energy Systems, Inc.)

2. TOWARDS SEQUENCING LONGER FRAGMENTS BY DNA TRAPPING ELECTROPHORESIS

Levy Ulanovsky, Guy Drouin, and Walter Gilbert; Harvard University.

The attachment of a neutral protein to the end of a single-stranded DNA molecule profoundly alters the mobility of the DNA on polyacrylamide gels. The combination of the observed phenomenon with field inversion electrophoresis increases the band separation severalfold in a desirable and controllable size range. This increased size separation may raise the efficiency of DNA sequencing by allowing the sequencing of much longer fragments, possibly up to several kb.

While short DNA molecules in complex with a neutral protein are retarded only a small amount under constant voltage, long molecules are retarded dramatically above a threshold size, 0.6 kb at 60 V/cm; at this voltage, molecules above a cut off about 1 kb will not enter the gel. Both the threshold and the cut off sizes increase as the voltage decreases. In field inversion electrophoresis, the longer the reverse pulse -- the larger the molecules that enter the gel.

These results are interpreted as trapping by the gel matrix of the protein attached to the DNA. The probability of release under constant voltage depends on the balance between the electric field and the thermal motion: the longer the DNA and the higher the voltage, the harder it is to untrap the molecules. Under field inversion conditions, the probability of untrapping increases with the duration and voltage of the reverse pulse and decreases with the DNA size.

3. MULTIPLEX DNA SEQUENCING WITH SANGER CHEMISTRY AND CHEMILUMINESCENT DETECTION.

Stephan Beck, Louis D'Angio, Theresa S. Dunne, Richard P. Hodge, Carol Kissinger, Thomas O'Keeffe, Thomas Templeman and Hubert Köster. MilliGen/Biosearch, 186 Middlesex Turnpike, Burlington, MA 01803, USA.

Multiplex DNA sequencing is a new and versatile sequencing strategy which is aimed mainly towards application in large scale sequencing projects. In its original form the multiplex sequencing technology is descibed for use with the Maxam and Gilbert chemistry (chemical degradation) and radioactive detection. Here we present recent modifications and improvemets to the multiplex technology which should make this technique easier and more convenient to use.

1) We have developed a non-radioactive DNA detection chemistry which is based on chemiluminescence. Alkaline phosphatase, either directly conjugated to the probe or bound to the probe via a biotin-streptavidin complex, is used as a reporter molecule for the hybridization. Upon exposure to the chemiluminogenic substrate the enzyme catalyzes a light reaction which can be detected by exposure to standard x-ray or Polaroid films.

2) We have developed protocols to simultaneously sequence multiple double stranded plasmid templates in a single reaction using the Sanger dideoxy chain termination chemistry. The routine use of the Taq DNA polymerase and 7-deaza Guanosine in the nucleotide mixes considerably reduce the inherent problems of sequencing double stranded DNA, such as secondary structures and compressions.

4. NEW SYNCHRONOUS LUMINESCENCE TECHNIQUE FOR IMPROVED DNA SEQUENCING(*), T. Vo-Dinh, R. W. Johnson, and H. F. Cheng, Advanced Monitoring Development Group, Health and Safety Research Division, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-6101 U.S.A.

This paper presents novel spectrochemical techniques developed to improve the speed and accuracy of DNA sequencing. Current technology using fluorescence-based techniques used straight-forward fluorescence detection with optical filters. Also, the conventional fixed-excitation fluorescence spectra of the dye-nucleotide conjugates used in current DNA sequencing techniques overlap substantially; this feature has previously led to misreading errors during sequencing utilizing conventional fluorescence detection because peaks corresponding to the presence of a single dye are detected in more than one channel. The synchronous luminescence (SL) technique developed at Oak Ridge National Laboratory was successfully applied to detect fluorescence signals from the four dye-conjugates, which are well resolved, thus leading to improved DNA sequence analysis. The results indicate that the SL technique allows improved spectral differentiation of separated species and therefore can shorten the sequencing time and at the same time improve the accuracy of DNA sequencing.

* Research sponsored by the Office of Health and Environmental Research, U.S. Department of Energy, under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

5. IMPROVED SEQUENCING OF DOUBLE-STRAND PLASMID TEMPLATES WITH MODIFIED TAQ DNA POLYMERASE. <u>David Thompson, Tom Burke, and Sandra Spurgeon</u>, Promega Corporation

Direct sequencing of double-strand plasmid templates is a rapid, convenient method of obtaining sequence data for cloned inserts. The use of Taq DNA polymerase for this purpose is advantageous because of the high temperature optimum, processivity and rapid rate of polymerization of this enzyme. However, with the native form of Taq DNA polymerase, we have observed high backgrounds that makes in interpretation of sequence data difficult when sequencing double-strand templates unless an end-labeled primer is used. This high background is probably due to the 5'-3' exonuclease activity associated with the native enzyme. Recently we have isolated a modified form of Taq DNA polymerase that has an apparent molecular weight of 80,000 compared to 85,000 for the native enzyme. This modified enzyme retains the high temperature optimum and processivity of the native enzyme, but is characterized by little to no background in sequencing reactions on double-strand templates. The use of the modified Taq DNA polymerase makes it possible to take full advantage of the desirable properties of Taq DNA polymerase on double-strand as well as single-strand templates.

MINIATURIZATION OF SEQUENCING BY HYBRIDIZATION (SBH): A NOVEL METHOD FOR GENOME SEQUENCING
<u>Crkvenjakov, R., Drmanac, R., Strezoska, Z., Labat, I.</u> Genetic Engineering Center, PO Box 794, 11000 Beograd, Yugoslavia

Human genome sequencing based on gel electrophoresis or recently proposed hybridization (Drmanac et al. GENOMICS (1989) 4:114) methods requires automated equipment on macro scale and can not be imagined as a routine procedure. Macro scale is mandated due to the requirements of robotic positioning of samples on predetermined coordinates and polymer separation in gels. However, determination of oligonucleotide contents of DNA which underlies SBH theoretically allows the micro scale processes with micro separated samples altogether comprising a macro scale reaction. It is possible to use the determination which clone/probe is on which random micro position instead of placing clone/probe on predefined macro position or volume. We propose the use of micro discrete particles (DPs) as vehicles for samples/probes. The recognition of specific association of a DP and a clone/probe is achieveable by premarking DPs and/or determining characteristics of clone/probe <u>in situ</u>. The most obvious ways of marking DPs are shape, size and color , or attaching to it a specific combination of known oligonucleotides. We offer two possibilities for human genome sequencing (Drmanac et al., manuscript in preparation). For direct SBH $1x10^7$ clones coming from 10 separate genome parts are bound to $1x10^6$ different DPs in as many macro reactions (or eventually in a single macro reaction). $1x10^3$ monolayers containing more than $1x10^7$ individual previously mixed DPs are each after DP identification hybridized with groups of 100 differently labeled octamers. To this end we have developed conditions for reliable short oligonucleotide hybridization. For inverse SBH $1x10^{7-9}$ different DPs are prepared each carrying a unique 12-15mer and unique combination of 20 of 40 marking oligos. No more than 5000 separate macro reactions are needed for their preparation. After 40 hybridizations with marker oligos to find association between 12-15mers and DPs in a monolayer, in 1-100 hybridizations with fragmented, end labeled human DNA data for sequencing are generated. The monolayer area covers at most 100 microscope slides.The data collection for both approaches needs automated image analysis giving speed of data bits acquisition of $1x10^6$/s. Finaly a substantional computing has a major role to keep track of information and generate sequence (see accompanying abstract). The described miniaturization concept and ensuing savings make human genome sequencing immediately feasible in a laboratory pending technological development.

7. **DIRECT SEQUENCING OF AMPLIFIED DNA FROM HIV INFECTED INDIVIDUALS**

D Amorese, T Norris, M Kirsch, and K Adler, Medical Products Department, E. I. du Pont de Nemours & Co., Inc., Wilmington, DE 19880-0328

The ability to use automated DNA sequencing to sequence amplified DNAs from HIV infected individuals can not only aid in early diagnosis of infection, but the data obtained can also be used to study the rate of mutation within a given gene. This may be helpful in identifying regions that should be targeted from a therapeutic or diagnostic perspective, as well as regions that mutate too rapidly and should, therefore, be avoided.

Techniques have been developed for direct sequencing of polymerase chain reaction (PCR) amplified DNAs on a fluorescence based, automated DNA sequencer, the Genesis (tm) 2000. A region of the HIV pol gene was amplified, from peripheral blood lymphocytes obtained from infected individuals, using three different approaches; standard symmetric, asymmetric, and biotinylated symmetric (one standard oligonucleotide primer and one biotinylated oligonucleotide primer[1]). The standard symmetric and asymmetric products were sequenced using traditional approaches. The biotinylated DNA was bound to streptavidin that had been immobilized on magnetic beads. The beads were washed and the non-biotinylated strand removed by sodium hydroxide treatment (this strand was recovered by ethanol precipitation). Both strands were then sequenced using the standard fluorescent dideoxy sequencing procedure for single stranded DNA.

This report describes the procedures in detail and compares the quality of the sequence obtained using biotinylated symmetric, standard symmetric, and asymmetric amplification products with amplifying and nested primers.

[1] Mitchell,L.G., and Merril,C.R. (1989) Anal. Biochem. 178.

• **TAQ THERMAL CYCLING FOR DIDEOXY SEQUENCING.**

Sandy M. Koepf, Leslie Johnston-Dow, James A. Fisher, Marianne Hane, Lincoln J. McBride, and Cheryl Heiner, Applied Biosystems, 850 Lincoln Centre Dr., Foster City, CA., 94404.

We have focused on the thermal stability of *Taq* polymerase and fluorescent detection to develop a DNA sequencing protocol. This protocol uses thermal cycling of the sequencing reactions and high reaction temperatures to generate a linear amplification of the extension products. Because of the thermal cycling, the number of pipeting manipulations are reduced and the amount of template needed is lower than that of standard sequencing protocols. After thermal cycling, the fluorescently-labeled extension products are detected on an Applied Biosystems Model 370A DNA Sequencer. This protocol has been used with single-stranded, double-stranded and PCR-generated templates with little alteration to the sequencing reactions between template types. We are able to read between 300 to 500 bases per sequencing reaction.

9. **Optimization of asymmetric polymerase chain reaction for rapid fluorescent DNA sequencing.**

Chia Chen, Richard K. Wilson, Ben Koop and Leroy Hood.
Division of Biology, 147-75, California Institute of Technology, Pasadena, CA 91125.

We will describe a high-throughput method for preparation of single-stranded template DNA suitable for sequence analysis using the fluorescent labeling chemistry. In this procedure, the asymmetric polymerase chain reaction (APCR) is employed to amplify recombinant plasmid or bacteriophage DNA directly from colonies or plaques. The use of amplification primers located at least 200 base pairs proximal to the site of sequencing primer annealing removes the need for extensive purification of the APCR product. Instead, the single-stranded product DNA is purified by a simple isopropanol precipitation step and then directly sequenced using fluorescent dye-labeled oligonucleotides. This method significantly reduces the time and labor required for template preparation and improves fluorescent DNA sequencing strategies by providing a much more uniform yield of single-stranded DNA. We have employed the APCR strategy to determine the nucleotide sequence of a 10,000 base pair segment of the murine T-cell receptor α / δ-chain locus (Wilson et al., in preparation.).

10. AUTOMATED APPROACHES TO DNA SEQUENCING

Bruce A. Roe, Elaine Mardis and Stephanie Pyle Department of Chemistry and Biochemistry, University of Oklahoma, Norman, Oklahoma 73019

If large scale DNA sequencing projects are to be realized, methods must be developed to isolate and manipulate purified template DNA which can reproducibly yield highly accurate DNA sequence data. We therefore now have developed automated procedures for simultaneously isolating 96 single-stranded M13 chimeric template DNAs in less than two hours, and for pipetting 24 simultaneous dideoxynucleotide sequencing reactions on a commercially available laboratory workstation. The DNA sequencing results obtained by either radioactive or fluorescent approaches are consistent with the premise that automating these portions of a DNA sequencing project can eliminate much of the tedium while improving the reproducibility of the DNA isolation and the dideoxynucleotide sequencing reaction pipetting steps. Implementation of automated procedures for these normally labor intensive steps can provide a reasonable approach to rapidly obtain large amounts of high quality, reproducible DNA sequence data.

11. AUTOMATION OF THE EXONUCLEASE III DIGESTION REACTION USING A BIOMEK 1000 AUTOMATED LABORATORY WORKSTATION. Anthony R. Kerlavage and J. Craig Venter Section of Receptor Biochemistry and Molecular Biology, LMCN, NINDS, NIH, Bethesda, MD 20892.

The exonuclease III digestion strategy for generating overlapping deletion clones is a powerful method for sequencing long segments of DNA which eliminates the problem of gaps in the sequence (1). The method is rapid, non-random and relies on the uniform rate of digestion of the enzyme. Automation of the digestion reaction and subsequent blunt-ending, ligation and transformation steps using the Biomek 1000 Automated Laboratory Workstation has provided a means of taking advantage of the uniformity of the reaction to generate deletion subclones of optimal length to provide the highest sequencing efficiency while maintaining an adequate degree of redundancy for a large scale genomic sequencing project. Both time and temperature of the reaction can be precisely controlled to provide deletions of any desired size. Inserts of up to 7kb can be processed with as few as 24 time points. The procedure performs all steps from digestion through transformation in a rapid, automated fashion in a 96-well format, allowing processing of a large number of samples per day.

(1) S. Henikoff, Gene 28:351 (1984)

12. RAPID DNA SEQUENCING USING THE MULTIWELL MICROTITRE PLATE SEQUENCING SYSTEM CONTAINING LYOPHILISED REAGENTS.

Peter J.Ellis, Ian R.Felix, Alison C.Sweet, Ronald H. Jackson and Jeremy N B Walker. Amersham International, White Lion Road, Amersham, Bucks, UK, HP7 9LL.

A rapid manual microtitre plate sequencing system has been developed which allows the sequencing reactions for 32-64 clones to be performed in one hour. The format requires an 8-channel pipette to transfer the DNA template into the first row of a microtitre plate to resuspend dried annealing buffer, followed by sequential transfer into rows containing dried reagents for each step in the reaction. Eight samples are processed using each microtitre plate. Under the conditions used the quality of the sequence ladder obtained is excellent and equivalent to that obtained before lyophilisation. The flexible format can be applied to a number of sequencing approaches including M13 dideoxy sequencing with universal or alternative primers, labelled primer sequencing and plasmid sequencing. Radioactive label is used in the form of (α-^{35}S)dATP dried in microtitre strip wells.

13. THE USE OF AUTOMATED, PREPARATIVE ELECTROPHORESIS TO FACILITATE SUBCLONING OF FRAGMENTS FOR SEQUENCE ANALYSIS.

W. Richard McCombie, David Iovannisci,* J. William Efcavitch* and J. Craig Venter. Section of Receptor Biochemistry and Molecular Biology, NINDS, NIH and Applied Biosystems, Inc.*

The subcloning of restriction fragments into sequencing vectors is an important part of ordered sequencing strategies. We have used an Applied Biosystems HPEC to perform preparative electrophoresis on various DNA samples. This device separates fragments electrophoretically, monitors the absorbance of the gel eluent and automatically fractionates the eluent into microcentrifuge tubes. We have used the HPEC to separate fragments of cosmid or lambda clones digested with one or a pair of restriction enzymes. Several gel concentrations were used. Fragments ranged in size from over 20kb to less than 1kb. Amounts of DNA ranged from 1 microgram to 10 micrograms per gel. The fragments could often be separated so that a single fragment was visible when the appropriate fraction was analysed on a standard agarose gel. In most other cases, while multiple fragments were present, there was substantial enrichment of one or a few bands in a given fraction. Modifications in current, fraction size and gel percentage should improve the purification possible with this device. The fractions could be used as a source of insert DNA in ligations to sequencing vectors. This should allow fractions collected by the HPEC to be transfered directly to other automated workstations with little human intervention.

14. AN AUTOMATED DNA SEQUENCE FILM READER, Steve Ferris, Bio-Rad Laboratories, Richmond, California 94804.

Bio-Rad Laboratories has developed an automated DNA autoradiogram scanner. It provides automatic and accurate base assignments from conventional Sanger dideoxy sequencing films. The film reader scans consecutive sections of the autoradiogram using a high resolution CCD detector. The digitized image data is then processed by an 80386 computer. Pattern recognition software finds the lanes, corrects for distortions and artifacts, then assigns each band as A, G, C, or T. Expert rules are employed in the final stage of analysis. Verification and editing are easily accomplished with a side by side comparison of the assigned sequence and gel image on the video monitor.

15. A DNA SEQUENCE FILM READING WORKSTATION. Murray R. Summers†, Mike Kelly†, Abe Coriat§, Ilya Ravkin§, Al Delfino§; (†) IntelliGenetics, Inc. and (§) Applied Imaging, Inc.

We describe a simple image acquisition device useful for processing conventional DNA sequencing autoradiograms. This MS-DOS based system using off-the-shelf hardware can acquire the image of sequencing films at varying resolutions, enhance the image to improve readability, correct the image for minor gel or electrophoretic anomalies, and decipher the pattern of bands to produce a DNA sequence. Images can be magnified to examine close band spacing. Assignment of individual residues can be completely interactive, or automatic. Once assigned, all residues are editable while the film image is displayed. Completed sequences can be saved to files in several formats, eliminating typographical and transcriptional errors. Images can be saved on the internal hard disk or on WORM disks, allowing several years worth of film data to be stored. Recalled images also contain sequence assignment information. Use of this device can reduce the amount of time required to construct the sequence from a single lane set by up to 90%, depending on the quality of the autoradiogram. Use of a system such as this is ideal in a high throughput, computerized sequencing facility.

16. RAPID TEMPLATE PURIFICATION METHODS FOR DNA SEQUENCING: CONVENTIONAL AND AUTOMATED DNA SEQUENCING OF SEPHAGLAS™ M13 MINIPREPS.
S. Duthie, D. Boyer, B. Walsh, T. Pham and H. Osterman, Pharmacia P-L Biochemicals Inc.,2202 North Bartlett Avenue, Milwaukee, Wisconsin 53202.

The Sephaglas™ M13 Miniprep procedure is designed for the rapid extraction and purification of single-stranded DNA from cultures of M13 phage using Sephaglas SS, a novel glass matrix from Pharmacia LKB. Sephaglas readily forms a suspension which is easy to handle, and it has well-defined DNA binding characteristics which allow the reproducible recovery of small amounts of DNA from minipreps. The matrix selectively binds DNA at a high concentration of chaotropic salt (sodium perchlorate) which denatures and removes proteins; DNA is then eluted in buffers of low ionic strength at sufficient concentration for direct use in DNA sequencing or mutagenesis. Many single-stranded templates can be prepared simultaneously in less than an hour without conventional extraction or precipitation.

M13 clones containing inserts of various sizes were prepared by the Sephaglas procedure and sequenced by either conventional methods or on the A.L.F. (Automated Laser Fluorescent) DNA Sequencer from Pharmacia LKB.

17\. EFFICIENCY CHARACTERIZATION OF A FLUORESCENCE BASED AUTOMATED DNA SEQUENCER

AUTHORS: F. H. Fraser, B. K. Burnett, J. P. Mizzer, A. M. Hochberg, and J. W. Eveleigh
E. I. Du Pont de Nemours & Co., Inc. Wilmington, DE 19898

Using a quality control database of 280 kb determined in August 1989, the overall efficiency of Du Pont's Genesis(tm) 2000 DNA Analysis System was characterized relative to elapsed time, labor time (both active and idle), and machine time per base pair determined at the 98% accuracy level. Results are compared to manual radio-isotopic sequencing, and an overall cost comparison is made using typical labor and reagent unit costs. Finally, a projection of the "plant capacity" necessary to accomplish just the sequencing phase of the human genome project using available automated fluorescent sequencers is made as a function of project time and sequencer accuracy level.

18\. FLUORESCENT DNA SEQUENCING USING DYE-LABELED NUCLEOTIDE TERMINATORS

Vergine Chakerian, Charles R. Connell, Steven Fung, Davis Hershey, Linda G. Lee and Sam L. Woo,
Applied Biosystems, Inc., 850 Lincoln Centre Drive, Foster City, CA 94404.

Fluorescence-based sequencing systems have had a major impact in advancing automated DNA sequencing technology in the past few years. Our research group has developed a set of four dye-labeled dideoxy nucleotide analogs that act as chain terminators in Sanger-type enzymatic sequencing. Termination and fluorescent labeling occur simultaneously in such a sequencing system. The set of four dye terminators is the result of the synthesis and testing of numerous nucleotide analogs. In our sequencing protocol all four dye terminators at micromolar concentrations are used in a single reaction mixture containing 1 ug of template DNA. The dye terminators were chosen and the system optimized for use with Taq DNA polymerase. Samples were run on an Applied Biosystems DNA Sequencer. Automatic base calling is highly accurate for both ss and ds templates to 450 and 300 nucleotides respectively.

19. TOWARDS LARGE SCALE DNA SEQUENCING. John A. Brumbaugh*, Lyle R. Middendorf**, Daniel L. Grone**, and Jerry L. Ruth***. *School of Biological Sciences, University of Nebraska, Lincoln, NE; **LI-COR, Inc. Lincoln, NE; ***Molecular Biosystems, San Diego, CA.

For the purposes of sequencing large amounts of DNA, an electrophoretic gel is scanned by a fluorescence microscope moving back and forth across the gel at a fixed vertical position. Each scan provides a horizontal input for the creation of a two-dimensional image. The vertical input is provided by the movement of fluorescently-marked bands of DNA molecules in the gel and as such bands cross the scanning location over time a picture is constructed line by line. The full two-dimensional appearance of the bands is displayed on a monitor. Utilization of two-dimensional image processing, either by the human brain or through computerized image analysis allows for the accurate sequencing of long clones of DNA. Information that is necessary for proper interpretation is not lost as in the case of comparing four one-dimensional graphs or curves.

Since mid-1986 over 200,000 bases have been sequenced with virtually error-free results for strand lengths of 400 bases and >99% accuracy for strand lengths of 600 bases. Present capabilities allow for the sequencing of about 4 bases per minute of operation per machine.

.0. A PHAGE T4 IN VITRO DNA PACKAGING SYSTEM FOR CLONING LONG DNA MOLECULES V. BASAVESWARA RAO* AND LINDSAY W. BLACK, Dept. of Biological Chemistry, Univ. MD. Medical School, 660 W. Redwood St., Baltimore, MD 21201. *Present Address: Dept. of Biology, The Catholic Univ. of America, Washington, D.C. 20064.

Phage T4 packages about 170 Kb of DNA that includes 2% terminal redundancy, in a strictly headful manner. Appropriate T4 mutations allow generalized T4 transducing particles at roughly the levels seen for phage P1. We have recently developed an *in vitro* T4 DNA packaging system using purified components. We are currently using this system to clone about 150 Kb size foreign DNA, and to construct genomic libraries of that size. Bacterial recipients used for transduction express lambda immunity (to restrict the growth of the T4rII packaging mutant phages used for extracts), and a phage recombinase to recircularize the injected recombinant DNA at a specific DNA sequence. We packaged a variety of concatemerized foreign DNAs such as the phage lambda DNA, cosmid DNA, and phage P1-pBR322 plasmid DNA, at efficiencies comparable to T4 DNA. We have cloned the 95 Kb *tryp*, and the 100 Kb *nad* fragments of *E. coli* into the P1-*lox* vectors developed by Dr. Nat Sternberg. While it is clear from the above data that the system can be used to clone long DNA molecules of any sequence, the major limitation at this stage to effective libraries construction appears to be the generation of intact 150 Kb size DNA fragments which can be manipulated for cloning.

21. GENOMIC SEQUENCE OF A 51 BASE PAIR REPEAT GENE OF PLASMODIUM FALCIPARUM: IMPLICATION OF SEQUENCING REPETITIVE DNA SEQUENCE OF HUMAN GENOME.

Jingdong Zhu and Michael R. Hollingdale. Biomedical Research Institute, 12111 Parklawn Drive, Rockville, Maryland, 20852, U.S.A.

In human, about 20% of the DNA is in repeated or closely related sequences, and these sequences can either be tandemly iterated (e.g., ribosomal DNAs) or dispersed throughout the chromosomes (e.g., retrotransposons and retrotransposon-like elements). The determination of these sequences may require more complicated and specific techniques. Some of current sequencing stratages (e.g., shotgun, PCA techniques) may not be suitable to define those with hundreds of tandem repeats. The genome size of parasitic protozoan Plasmodium is about 107 bp but consists of up to 10% repetitive sequences. We have cloned a 5.2 kp P. falciparum genomic fragment by oligonucleotide hybridization to a liver stage antigen gene, part of which has recently been reported (Guerin-Marchand et al. 1987, Nature 329:164). Using the principle of ordered deletion method (Henikoff 1987, Methods in Enzy. 155:156), we were able to simplify the steps before sequencing gel electrophoresis and finish raw sequencing of about 5kb DNA fragment in a week. This sequence of this fragment contains at least 23 copies of a 51-bp major tandem repeat interrupted by 3 minor repeats. The techniques have the implications of sequencing tandem repeats in human genome.

22. RAPID COSMID FINGERPRINTING AND CONTIG ASSEMBLY. E. Branscomb, P. de Jong, and A.V. Carrano. Biomedical Sciences Division, Lawrence Livermore National Laboratory, Livermore, CA.

High resolution gel electrophoresis provides a simple and versatile method for DNA fingerprinting and the creation of contigs or sets of overlapping genomic clones. We have developed an approach to cosmid contig construction that: 1) uses a robotic system to label restriction fragments from cosmids with fluorochromes; 2) uses an automated DNA sequencer to capture fragment mobility data in a multiplex mode (i.e. three cosmids and a size standard in each lane); 3) processes the mobility data to determine fragment length and provide a statistical measure of overlap among cosmids; and 4) displays the contigs and underlying cosmids for operator interaction and access to a database. We have applied these methods to construct a cosmid contig map for a 600 kbp YAC clone from chromosome 14 and are currently analyzing cosmids to construct contigs for all of chromosome 19. Throughput rate is currently about 42 cosmids per day per machine but 96 cosmids per day is achievable. Resolution of fragment size is to within 1–1.5 bases over the range of 29–462 bases for which data are captured. The 1510 chromosome 19 cosmids analyzed to-date form into 206 contigs with an average contig length of 3.2 cosmids. Several of these contigs have been located to the chromosome by fluorescence *in situ* hybridization and also mapped to known genes. The "minimal" spanning sets of cosmids provide unique starting material for genome sequencing. This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.

23. METHOD AND STRATEGY FOR DIRECTLY SEQUENCING COSMID INSERTS: SEQUENCING THE HUMAN T-CELL RECEPTOR β-GENE FAMILY [1]Jerry L. Slightom, [1]David R. Siemieniak, [1]Leang C. Sieu, [2]Leroy Hood, and [3]Morris Goodman [1]Molecular Biology Unit 7242, The Upjohn Company, Kalamazoo, MI 49007; [2]Division of Biology, California Institute of Technology, Pasadena, CA 91125; [3]Department of Anatomy and Cell Biology, Wayne State University, Detroit, MI 48201.

We are presently developing DNA sequencing technology which avoids the time consuming steps of subcloning, subclone mapping and multiple sequencing of identical clones. Our method involves the use of both chemical and enzymatic sequencing procedures to directly obtain the sequence of cosmid inserts in a directional manner. Chemical sequencing is used to sequence small, unmapped, end-labeled fragments obtained from the sequential use of six base pair cutting enzymes. Generally three to four unmapped DNA fragments are chemically sequenced for each cosmid insert, followed by obtaining oligomer primers which correspond to the 3' and 5' ends of each sequenced region. The location of these "random" sequence regions or contigs can be determined by PCR mapping. Enzymatic chain termination sequencing of double-stranded cosmid DNAs are accomplished by using T7 polymerase, and sequences are run on 1 meter gels which allow for reads in excess of 600 bp. Sequence contigs are continued until closure is achieved. This technique is being used to sequence the 84 kb region of the human T-cell receptor β-gene family contained with overlapping cosmid clones H7.1, H12.18, and H130.1 (Lai et al., 1988 Nature 331, 543-546).

24. NUCLEOTIDE SEQUENCE DETERMINATION OF THE GALAGO 45 KB β-TYPE GLOBIN GENE CLUSTER [1]Danilo A. Tagle, [2]Philip Benson, [2]David Fitch, [2]Michael Stanhope, [3]Jerry L. Slightom, [1,2]Morris Goodman. Depts. of [1]Molecular Biology & Genetics, [2]Anatomy, Wayne State Univ., Detroit, MI 48201, [3]Div. of Molecular Biology, The Upjohn Company, Kalamazoo, MI 49007.

The nucleotide sequences of the galago β-like globin gene cluster spanning 45 kb was determined using chemical and double-stranded dideoxy sequencing methods. The latter is quicker and the synthetic oligomers used can be applied in PCR for vertical expansion of nucleotide sequence data from different species to detect phylogenetic footprints (i.e. evolutionarily conserved sequences that are candidate cis-regulatory elements). As expected, the exons showed a slow divergence rate, however, the non-synonymous mutation rate is noticeably slower in the embryonic ε gene than in the adult β globin gene. The rate of DNA evolution of $4.2-5.5 \times 10^{-9}$ substitutions/site/year for non-coding regions (flanking and pseudogene sequences) approximates that observed for synonymous sites thus indicating that these regions are evolving under little selective constraints. Open reading frames (ORF) were found not only in the known protein-coding genes but also in L1 repeat elements. Thus attempts to find genes using ORFs in a large scale sequence may be misleading*. Dot plot comparisons at a match criterion of 23 in a window of 30 showed extensive alignments of orthologous genic and flanking human sequences with galago but limited alignments with rabbit or mouse.

*see poster by Moore et al. "DNA Exon Boundaries Detected by Evolutionary Conservation".

IAE000507878

25. **BOVINE IMMUNODEFICIENCY-LIKE VIRUS: NUCLEOTIDE SEQUENCE ANALYSIS AND EVOLUTIONARY COMPARISON TO OTHER RETROVIRUSES** -

Kevin J. Garvey, M. Steven Oberste, and Matthew A. Gonda, Laboratory of Cell and Molecular Structure, Program Resources, Inc., National Cancer Institute-Frederick Cancer Research Facility, Frederick, MD 21701.

The nucleotide sequences of two biologically active, proviral clones (BIV 106 and 127) of the bovine immunodeficiency-like virus, obtained from a single isolation, have been determined. The genome of BIV 106 (8,391 nt) has suffered several deletions relative to BIV 127 (8,482 nt); numerous substitutions have also occurred. Despite this, the organization of the two clones is identical and most similar to that of the lentivirus subfamily of retroviruses (i.e., LTR-*gag-pol*-central region-*env*-LTR). There are five ORF's in the central region; three have been tentatively identified as the regulatory genes, *vif*, *tat*, and *rev*, present in other lentiviruses and two are unique to BIV. Comparisons of the translations of the structural genes of BIV with those of other lentiviruses identified regions of conserved amino acid sequence which permitted us to predict the mature structural gene products. Our present results suggest that individual genomes of BIV from a single isolate are quite variable and the greatest variability occurs in the outer membrane protein, a trait it shares with HIV.

26. SOFTWARE FOR TRANSFER AND GRAPHICAL EDITING OF DATA FROM AUTOMATED DNA SEQUENCERS.

James Stein, Janice McCombie*, Robert Jones, Jeannine Gocayne*, W. Richard McCombie*, Anthony Kerlavage*, and J. Craig Venter*. Applied Biosystems, Inc. and Section of Receptor Biochemistry and Molecular Biology, NINDS, NIH*.

One of the major advantages of automated DNA sequencers is their ability to generate machine readable DNA sequence data. We are developing a series of software tools and programs to capitalize on that advantage. These programs will reformat sequence data from the ABI 370A DNA sequencers for use by the MBIR (Baylor) or UWGCG (Wisconsin) analysis packages. Included in these tools is a multiple sequence editor which can display along with each sequence, its chromatogram (the graphical representation of the data produced by the sequencers). We are developing a conversion program that will allow a "contig" produced by the MBIR Sequence Assembly Manager ("SAM") to be read into the graphical sequence editor. Being able to display easily the graphical data for each sequence will greatly facilitate the resolution of discrepancies in the contig. Another conversion program will convert the layout file produced by the editor into sequences that can be put back into the "SAM" project so that additional sequences can be merged into the edited contig, if necessary.

27. ALGORITHM FOR SEQUENCE GENERATION FROM K-TUPLE WORD CONTENT

Drmanac, R., Labat, I., Crkvenjakov, R. Genetic Engineering Center, Belgrade

Any text, as well as nucleotide sequence can be represented as a set of the overlaping k-tuple words, similary to the methods applied in the most efficient sequence comparison algorithms used today. Our algorithm uses intrinsic nucleotide sequence informatics to regenerate the original sequence without k-tuple position and and frequency information inherent to the former algorithms. K-tuples are ordered by maximal overlaping up to the moment when none, or two or more k-tuples overlap with last one attached. Further ordering is ambiguous. A primary subfragment (SF') is thus defined. Number of the heuristic methods developed repairs and unambiguously connects SF' into the real subfragments (SF). Number and length dispersion of the SFs as sequence informatics entities depends on length and simplicity of the sequence as well as length of k-tuples and extent of mistakes in the set. The other part of the algorithm enables regeneration of the sequence of the length of the human genome fragmented in the suggested manner (Drmanac et al., GENOMICS '89,4,114). It consists of several k-tuples sets and SFs manipulations on different, overlaping sequence fragments. Our software is applied on the IBM PC/AT compatible. In simulation experiment on the 50kb sequence, complete k-tuples sets (k=8 to 12, depending on GC content) of the consecutive nucleotide sequence fragments up to 900bp were handled. In over 91% of analyzed fragments the complete sequences were regenerated. In remaining cases, sequences were regenerated to the level of several (below 15) SFs. Also, 10% of false negative k-tuples in the set makes no problem in most of analyzed sequences. Further improvements of algorithm are needed (and suggested) for complete regeneration of some specific sequences, and for using sets with more false positive and false negative k-tuples.

28. THE USE OF ARTIFICIAL INTELLIGENCE METHODS IN AUTOMATED DNA SEQUENCE ANALYSIS SOFTWARE.

T. Allen, C. Shearer-Cooper, and G. Page. EG&G Biomolecular, Natick, MA.

The accurate interpretation of DNA sequence data collected by instrumentation has been and will continue to be a significant problem in automated sequence analysis. As a data type, DNA sequence exhibits great variability. The variations occur in signal strength from peak to peak, interpeak spacing, signal to noise ratio, peak resolution, and the distribution of signal in the data set. At the same time, the requirements for accuracy in interpretation are high. Current expectations demand accuracies of 99% or greater. But even this may be inadequate for very large scale sequencing projects.

Certain aspects of DNA sequence analysis lend themselves to artificial intelligence-based methods of processing. Software which exhibits artificial intelligence possesses some knowledge about the problem in hand and applies this knowledge to arrive at a solution. Areas in which these techniques have proven to be applicable include resolution of conflicts, identification of compressions, measurement of interpeak spacing, and determination of the point in each sequence at which further analysis becomes too inaccurate.

29. GENLANG: A COMPUTATIONAL LINGUISTIC SYSTEM FOR SEQUENCE DESCRIPTION, SEARCH AND ANALYSIS

Michiel O. Noordewier (Computer Science Department, University of Wisconsin)
David B. Searls (Paoli Research Center, Unisys Corporation)

We have previously introduced an approach to biological sequence analysis founded on principles and techniques from computational linguistics [Searls, AAAI-88, 386-391]. By describing elements of sequences with formal structural specifications known as phrase structure grammars, it is possible to search and analyze those sequences using well-known parsing algorithms. Such search produces a hierarchical description of the sequence called a parse tree. We have created a flexible parsing system for sequences using the formalism of Definite Clause Grammars (DCGs) and the logic programming language Prolog. Examples are given of grammars and parses for: (1) low-level features of the HIV genome; (2) high-level descriptions of globin genes that are able to accurately parse the entire beta-globin gene family from a 73kb region; (3) gene grammars with relaxed constraints that predict known splicing variants resulting from beta-thalassemia mutations; and (4) motif grammars that are able to recognize patterns of putative secondary structure features in polypeptides, in a hierarchical fashion similar to Ariadne [Lathrop et al, CACM 30, 909-921].

30. DNA EXON BOUNDARIES DETECTED BY EVOLUTIONARY CONSERVATION. G. William Moore, Stephen Williams, Jerry Slightom, Danilo Tagle, David Fitch, and Morris Goodman. Baltimore Veterans' Administration Medical Center and Wayne State University School of Medicine.

It is important to predict boundaries between exons and non-coding segments of the genome. Current approaches predict exon boundaries by signal or by content. Non-synonymous (amino acid changing) DNA base substitutions are evolutionarily suppressed in exons. We examined beta-globin, gamma-globin, and epsilon-globin sequences for human and Galago crassicaudatus, using the public-domain CONSOFT (conservation-software) package to compare non-synonymous base substitution rates in 2814, 2100, and 1996 consecutive bases from beta-globin, gamma-globin, and epsilon-globin. For 5' untranslated, intron 1, intron 2, 3' untranslated, and 3' flanking regions: 25%, 14%, 17%, 25%, 25%, and 21% in beta-globin; 16%, 13%, 25%, 25%, 22%, and 23% in gamma-globin; 13%, 18%, 12%, 22%, 24%, and 26% in epsilon-globin. For exons: 7%, 4%, and 4% in beta-globin; 8%, 5%, and 10% in gamma-globin; 8%, 4%, and 2% in epsilon-globin. Exons have fewer non-synonymous base substitutions than non-coding DNA, and thus may be located statistically by their relative scarcity of such substitutions. Exon boundary detection by conservation is a useful supplement to detection methods by signal and by content.

31. ESSENTIAL ROLE OF THE PROTEIN SEQUENCE DATABASE IN GENOME ANALYSIS

W.C. Barker, D.G. George, and L.T. Hunt, Protein Identification Resource, National Biomedical Research Foundation

The Protein Sequence Database (PSDB) is an essential component in genome sequencing projects. The 3 kinds of sequence databases — nucleic acid, protein, and X-ray crystallographic — form a continuum of complementary information, each having its own characteristic organization, information content, and special features. Sequences in the PSDB are organized by degree of similarity between the proteins, which has proved to be a powerful concept for investigations of protein relationships and functions. As sequencing proteins yields information quite distinct from that inferred from corresponding nucleotide coding regions, the PSDB contains information not available in the nucleic acid sequence databases that must be used, often in conjunction with nucleic acid sequences, to interpret correctly or confirm dimensions of protein coding regions, exon boundaries and their correlation with protein domains, operation of genetic mechanisms such as exon shuffling and alternate splicing, posttranslational modifications, and genetic variation in populations. As sequencing of various genomes proceeds, searching databases will aid in identifying the encoded proteins and their normal functions. Searching the PSDB to find closely or distantly related sequences has many advantages: comprehensiveness (for many proteins, a nucleotide sequence has not been determined), efficiency and speed, and sensitivity. Nucleic acid database searches can identify only highly similar sequences, whereas techniques developed for protein database searching allow very distant protein relationships to be discovered; the latter are often of greatest interest. The PSDB has been cited as being instrumental in the identification of a number of protein sequence similarities of medical significance in the past few years, incuding proteins involved in Alzheimer's disease, AIDS virus infection, demyelinating diseases, cystic fibrosis, cellular control mechanisms and oncogene function, immune system components, and complex endocrine-activated pathways. Supported by NIH grant RR01821.

32. DEVELOPING METHODOLOGY FOR CLONE CHARACTERIZATION IN A REPOSITORY OF DNA CLONES

D.A. Aebig, A.S. Durkin, D.L. Perino, J. Tomlin, D. Maglott, W.C. Nierman, American Type Culture Collection

An important function of the American Type Culture Collection (ATCC) is to provide the scientific community with well-characterized, cloned DNA segments from a variety of taxa to support gene structure and function studies and genomic mapping efforts. Having established and operated for several years a repository of human and mouse clones and libraries, the ATCC has begun a project to evaluate methods to maintain a repository on the scale of the human genome project. A model system for processing clones and information for a complete, but smaller, genome has been generated by using a set of about 1000 minimally overlapping genomic clones from Saccharomyces cerevisiae provided by Dr. Maynard Olson at Washington University. Procedures being evaluated in the laboratory include automation of sample preparation and handling, use of the ABI sequencer for restriction fragment analysis, and preservation of clones both for archival storage and for distribution in phage, cosmid and yeast artificial chromosome vectors. Computerized methods to manage information about Repository materials are being developed not only at the laboratory notebook level, but also by integrating genetic and physical maps.

Supported by the DOE DE-FG05-88ER60686 and the NIH 1R01 GM4293401